Normalization of TCGA File Submission and Distribution: New Submission Process for BCRs

Executive Overview

BCRs are now able to upload partial revisions of archives. The standard serial archives are split by data Level. There is currently only Level 1 data for BCRs:

- Level 1 (*e.g.* intgen.org_GBM.bio.Level_1.1.0.0.tar.gz)

Data Levels are determined using the Data Type-Data Level Matrix (DTDLM; https://gforge.nci.nih.gov/frs/?group_id=265). Archives contain only files that have the data Level of the archive. If you have a data type that is not listed in the DTDLM, please contact the DCC.

All archives, revised or not, must *always* contain a MANIFEST.txt file. The MANIFEST.txt lists all the names of the files to be included in the final (current or revised) archive and those files MD5 hashes. The format of the MANIFEST.txt should be consistent with MD5sum output: MD5_hash <space> filename. For example:

```
b1c6647b5f3b555a13de8af48b99920f broad.mit.edu_GBM.ABI.1.maf
abceb6390bb9c9dfe79695404268ad70 broad.mit.edu_GBM.ABI.1.vcf
```

A straightforward method of creating a MD5 MANIFEST.txt file is to use md5sum (*e.g.* `md5sum * > MANIFEST.txt`).

A README_DCC.txt for each archive will be composed by the DCC to include standardized text that describes the center's experiment and how to relate all the parts. A CHANGES_DCC.txt document that describes which files have changed between revisions by comparing MANIFEST MD5s will also be created by the DCC. Optional text documents may be included by the center in each archive that further describes anything about the archive files.

Archives should be compressed using tar and gzip. All files are expected to be contained in one flat directory when un-archived; that is, no subdirectories in an archive. A summary of the major modifications required and the reasoning behind those modifications is provided in .

Table 1 - Reasons for Modifications

Modification	*Reasoning*
Splitting data into archive by levels	Centers need only upload the files required for revision and users only need to download the data levels they are interested in.
New MANIFEST.txt format with MD5s	Guarantees the intended addition, replacement, or deletion of files is accurate

DCC creation of README.txt	Provides consistent and useful information to the end user

New Submissions

A new submission is composed of any Level archives where data is available that are not a revised archive. Each archive contains a MANIFEST.txt that lists all files in the archive and those files MD5 hash. All archives are compressed using tar and gzip and MD5-hash files are created for each compressed archive.

Revised Submissions

All new or revised files should be submitted in the context of the archives created above. That is, adding new files or revisions of current files, should be submitted as revisions of your Level 1 archives.

To create a revised archive by *adding files to*- or *replacing files in* a distributed archive, transfer only the files that are to be added or replaced, and each changed archive's MANIFEST.txt. To create a revised archive by removing a file from the currently distributed archive, transfer only the changed archive's MANIFEST.txt.

The process is as follows:

1. Create a new Level archive directory that has an incremented revision number in the archive name (*e.g.* intgen.org_GBM.bio.Level_1.1.0.0.tar.gz to intgen.org_GBM.bio.Level_1.1.1.0.tar.gz).
2. If adding or replacing files, add the new files to the archive in Step 1.
3. Create a MD5 MANIFEST.txt for the Level archive that contains a complete list of *all* the files that the revised Level archive *should* contain by copying the MD5 MANIFEST from the previous version of the archive and then modifying it. For example,
    a. If *adding* files, the MANIFEST.txt would be the same as the previous revision's MANIFEST.txt except for the added files' MD5 hashes and names.
    b. If *replacing* files, the MANIFEST.txt would be the same as the previous revision's MANIFEST.txt except that the replaced files would have a new MD5 hashes.
    c. If *removing* files, the MANIFEST.txt would be the same as the previous revision's MANIFEST.txt except that the files to be removed and their MD5 hashes would be deleted.
    d. If *keeping files as they are*, the MANIFEST.txt would be the same as previous revision's MANIFEST.txt with no change to the file names or their MD5 hashes, however those files are not included in the archive.
4. Compress the Level archive using tar/gzip
5. Create an MD5 for the compressed archive in Step 4

6. Run the DCC Validator on the revised archives. NOTE: The local version of the Validator will only verify the locally available files, it will not check for files that will be merged when submitted. Those files will be checked at time of transfer to the DCC.
7. If the revised archives pass validation, then transfer them to the DCC.

Submission Failures

A submission will fail if:

1. A MD5 hash is incorrect for any compressed archive.
2. A MD5 hash is incorrect for any file contained within an archive.
3. There is a mismatch in the files listed in the archive MANIFEST.txt.
4. A file listed in the archive MANIFEST lists a file that does not appear in the submitted archive revision, then the system checks the previous archive revision to see if it is there. If the file does not exist, then the submission fails.
5. If a MANIFEST.txt is not included in an archive.
6. Standard archive validation fails.